WHAT IS CLAIMED IS:

1. Method of encoding linguistic frequency data, the method comprising:

identifying a plurality of sets of character strings in a source text, each set comprising at least a first and a second character string,

for each set, obtaining frequency data indicative of the frequency of the respective set in the source text,

for each character string that is a first character string in at least one of the sets, assigning a memory position in a first memory array to the respective character string and storing at said memory position the frequency data of each set comprising the respective character string as first character string, and

for each character string that is a second character string in at least one of the sets, assigning a memory position in a second memory array to the respective character string and storing at said memory position, for each set comprising the respective character string as second character string, a pointer pointing to a memory position in the first memory array assigned to the corresponding first character string of the respective set and having stored the frequency data of the respective set.

2. The method of claim 1 wherein each set of character strings further comprises a third character string, and the method further comprises:

for each character string that is a third character string in at least one of the sets, assigning a memory position in a third memory array to the respective character string and storing at said memory position, for each set comprising the respective character string as third character string, a pointer pointing to a memory position in the second memory array assigned to the corresponding second character string of the respective set and having stored a pointer pointing to the frequency data of the respective set.

3. The method of claim 1 wherein each character string is a word of a natural language.

4. The method of claim 1 wherein each set of character strings comprising *n* character strings, *n* being an integer number greater than one, each set being an n-gram.

5. The method of claim 4 wherein *n* is equal to 3, the n-grams being trigrams.

6. The method of claim 1 wherein said frequency data indicative of the frequency of the respective set in the source text includes the number of occurrences of the respective set in the source text.

7. The method of claim 1 wherein said frequency data indicative of the frequency of the respective set in the source text includes weight numbers of a maximum entropy model.

8. The method of claim 1, further comprising:

mapping each character string occurring in the source text to a numeric identifier identifying the character string, by operating a finite-state machine.

9. The method of claim 8, further comprising:

storing the memory positions assigned to the respective first character strings in a first positional array,

storing the memory positions assigned to the respective second character strings in a second positional array, and

using the numeric identifiers to query the positional arrays for assigning the memory positions to the character strings.

10. The method of claim 1, further comprising:

accessing a hash-table for assigning a numeric identifier to each character string occurring in the source text.

11. The method of claim 10, further comprising:

storing the memory positions assigned to the respective first character strings in a first positional array,

storing the memory positions assigned to the respective second character strings in a second positional array, and

using the numeric identifiers to query the positional arrays for assigning the memory positions to the character strings.

12. The method of claim 1, further comprising:

in the second memory array, sorting the pointers relating to the same second character string, with respect to the memory positions of the first memory array to which the pointers point.

13. The method of claim 1 wherein the pointers are stored in compressed form.

14. Method of accessing encoded linguistic frequency data for retrieving the frequency of a search key in a text, the search key comprising a first and a second search string, the encoded data being stored in a first memory array storing frequency data and a second memory array storing pointers to the first memory array, the frequency data being indicative of the frequencies of character sets in a source text, the character sets each including at least two character strings, the method comprising:

identifying a region in the first memory array that is assigned to the first search string,

identifying a region in the second memory array that is assigned to the second search string,

identifying a pointer stored in the region of the second memory array, pointing to a memory position within the region of the first memory array, and

reading the frequency data stored at said memory position.

15. The method of claim 14 wherein the search key further comprises a third search string and the encoded data is further stored in a third memory array storing pointers to the second memory array, wherein the method further comprises:

identifying a region in the third memory array that is assigned to the third search string,

identifying a pointer stored in the region of the third memory array, pointing to a memory position within the region of the second memory array, and

tracing the pointer stored in the region of the third memory array back until the region of the first memory array is reached.

16. The method of claim 14 wherein each character string is a word of a natural language.

17. The method of claim 14 wherein each set of character strings comprising *n* character strings, *n* being an integer number greater than one, each set being an n-gram.

18. The method of claim 14 wherein *n* is equal to 3, the n-grams being trigrams.

19. The method of claim 14 wherein said frequency data indicative of the frequency of the respective set in the source text includes the number of occurrences of the respective set in the source text.

20. The method of claim 14 wherein said frequency data indicative of the frequency of the respective set in the source text includes weight numbers of a maximum entropy model.

21. The method of claim 14 wherein identifying a pointer includes performing a binary search within the second memory array.

22. The method of claim 14 wherein identifying a pointer includes identifying a sub-interval in the region of the second memory array, the sub-interval including at least two pointers pointing to a memory position within the region of the first memory array.

23. The method of claim 14, wherein identifying a pointer includes performing a first binary search for a set of pairs of strings where the first string in each pair matches the first search string, performing a second binary search for a set of pairs of strings where the second string in each pair matches the second search string, and calculating an intersection of both sets.

24. The method of claim 14, further comprising:

if the character sets in the source text comprise more character strings than the search key, arranging the search strings in the search key such that the potentially missing search strings appear first.

25. A system for encoding linguistic frequency data, comprising:

a processing unit for identifying a plurality of sets of character strings in a source text, each set comprising at least a first and a second character string, and, for each set, obtaining frequency data indicative of the frequency of the respective set in the source text, and

an encoder that, for each character string that is a first character string in at least one of the sets, assigns a memory position in a first memory array to the respective character string and stores at said memory position the frequency data of each set comprising the respective character string as first character string, and that, for each character string that is a second character string in at least one of the sets, assigns a memory position in a second memory array to the respective character string and stores at said memory position, for each set comprising the respective character string as second character string, a pointer pointing to a memory position in the first memory array assigned to the corresponding first character string of the respective set and having stored the frequency data of the respective set.

26. A system for accessing encoded linguistic frequency data for retrieving the frequency of a search key in a text, the search key comprising a first and a second search string, the encoded data being stored in a first memory array storing frequency data and a second memory array storing pointers to the first memory array, the frequency data being indicative of the frequencies of character sets in a source text, the character sets each including at least two character strings, the system comprising:

an input device for inputting the search key, and

a search engine for identifying a region in the first memory array that is assigned to the first search string, identifying a region in the second memory array that is assigned to the second search string, identifying a pointer stored in the region of the second memory array, the pointer pointing to a memory position within the region of the first memory array, and reading the frequency data stored at said memory position.